Publishing DisGeNET as Nanopublications

Editor(s): Boyan Brodaric, Geological Survey of Canada, Canada.
Open review(s): Amrapali Zaveri, AKSW, University of Leipzig, Germany; Eleni Mina, Leiden University, The Netherlands.

Núria Queralt-Rosinach[a], Tobias Kuhn[b], Christine Chichester[c], Michel Dumontier[d], Ferran Sanz[a], and Laura I. Furlong[a,*]

[a]*IBI group, Research Programme on Biomedical Informatics (GRIB), IMIM, DCEXS, Universitat Pompeu Fabra, Barcelona, Spain*
[b]*Department of Humanities, Social and Political Sciences, ETH Zurich, Switzerland*
[c]*CALIPHO group, Swiss Institute of Bioinformatics, CMU rue Michel Servet 1, 1211 Geneva 4, Switzerland*
[d]*Stanford Center for Biomedical Informatics Research, Stanford University, USA*

Abstract. The increasing and unprecedented publication rate in the biomedical field is a major bottleneck for knowledge discovery in the Life Sciences. The manual curation of facts from published scientific papers is slow and inefficient, and therefore new approaches are needed that can enable the automatic, scalable and reliable extraction of assertions. While the publication of scientific assertions and datasets on the Semantic Web is gaining traction, it also creates new challenges such as the proper representation of provenance and versioning. Here, we address these issues and describe our efforts to represent the DisGeNET database of human gene-disease associations as permanent, immutable, and provenance rich digital objects called nanopublications. Our nanopublications are the first instance of a Linked Data model that ensures stable interlinking of the assertion and its metadata by Trusty URIs. As DisGeNET integrates manually curated as well as text-mined data of different origins, the semantic description of the evidence for each assertion is important to provide trust and allow evidence-based hypothesis generation. Here, we describe our steps to ensure high quality and demonstrate the utility of linking our data to other datasets on the emerging Semantic Web.

Keywords: gene-disease associations, linked data, nanopublication, provenance, trusty URIs

1. Introduction

To obtain a deeper understanding of the molecular mechanisms of diseases and to support drug development and healthcare research, biomedical researchers need to explore the current knowledge on the complex relationships between genes, proteins, gene variants, pathways, drugs, phenotypes and environmental factors. Scientific knowledge is mainly communicated and gathered as scholarly publications. To ease the access and exploitation of this knowledge, one of the main strategies is to manually extract and curate biomedical statements from the literature and structure them in databases. Due to the increasing size of literature repositories and the number of different dispersed and isolated databases, there are many efforts devoted to efficiently extract and provide the most up-to-date data in a way that can be integrated with existing datasets to facilitate knowledge discovery. These efforts include a) text mining approaches aimed at extracting relationships between biomedical entities from the literature [10], b) community-driven publication approaches based on wiki systems [6-7, 22], c) the publication of existing databases to the Linked Open Data cloud (LOD)[1] (UniProt [24], DisGeNET [13]), and d) ambitious projects like the publication of entire Linked Data networks such as Bio2RDF [3] and Linked Life Data[2].

DisGeNET is a discovery platform developed in the Integrative Biomedical Informatics (IBI) group designed to enable research on the genetic basis of the pathophysiology of diseases. The platform offers one of the most comprehensive collections of knowledge on human gene-disease associations (GDAs) integrating over 380,000 associations between more than 16,000 genes and 13,000 diseases covering all disease areas. These GDAs are collected from seven different public databases, which include human and animal model expert-curated databases. DisGeNET also includes GDAs extracted from MEDLINE by the BeFree [2] NLP-based approach. All these data are integrated, harmonized and made accessible for exploration and analysis through a Web interface, a Cytoscape plugin [1], and as an RDF linked dataset [17] with an open license.

The Semantic Web enables data integration and interoperability, but new challenges emerge when the data are used for the identification and evaluation of scientific hypotheses. These challenges include the tracking of provenance to understand the basis of an assertion and its relation to existing evidence, and the creation of unambiguous references to immutable scientific assertions. To overcome these issues we publish our DisGeNET GDAs as a new linked RDF dataset using the emerging nanopublication approach[3] and the Trusty URI technique [23]. The nanopublication approach is a community-driven effort that proposes a publishing model based on minimal scientific statements along with their provenance and associated context. A nanopublication is formally defined using Semantic Web standards - the W3C's Resource Description Framework (RDF) and the Web Ontology Language (OWL) - and the nanopublication schema[4]. Specifically, it consists of three named RDF graphs (excluding the head graph) representing the assertion, its provenance and metadata about the nanopublication itself using resolvable Uniform Resource Identifiers (URIs). Consequently, standard RDF technologies, such as triple stores and SPARQL query engines, can be used to retrieve and analyze nanopublications through the Web, and to support automatic reasoning.

The linked dataset presented here is the first database published as nanopublications using Trusty URIs. Trusty URIs use cryptographic hash values to generate unique and stable identifiers based on their content. This makes digital artifacts identified using Trusty URIs permanent, immutable, and verifiable. Converting a dataset into nanopublications with Trusty URIs enables users and software agents to trace and interpret how a statement was produced, ensuring reproducibility, reliability, and enhancing citation. Another advantage is that both the scientific assertions and their metadata are interoperable, shareable, reusable, and supports discoverability by provenance-aware applications.

In this paper, we present the DisGeNET Nanopublications linked dataset as an alternative way to disseminate the information contained in DisGeNET. The conversion into nanopublications presented in this paper aims to extend and complement the capabilities of the existing DisGeNET dataset in RDF. The goal is to foster the publication and discoverability of these assertions, to support the automated aggregation of their evidence levels, and to support the generation of evidence-based new hypotheses in the biomedical field.

2. Gene-Disease Associations in DisGeNET

In order to cover different aspects of gene-disease relations, DisGeNET GDA content is extracted from various types of sources ranging from structured databases on human and animal models to unstructured scientific literature. DisGeNET provides evidence-based classifications of the data according to the level and type of curation in the original databases that enable users to rapidly assess the quality of the specific GDA. The DisGeNET evidence classes are: "CURATED" for human GDAs that are reviewed by experts, "PREDICTED" for human GDAs inferred from the GDA of an animal model that was reviewed by an expert, and "LITERATURE" for human GDAs that were automatically extracted from the literature by text mining methods (see DisGeNET coverage in Table 1). It is important to point out that DisGeNET not

[1] http://lod-cloud.net/

[2] http://linkedlifedata.com/

[3] http://nanopub.org/

[4] http://nanopub.org/guidelines/working_draft/

only aggregates GDA statements from different sources, but integrates them in a uniform way accompanied with contextual annotation. Specifically, the provenance and evidence are well described and fine-grained for each statement in order to keep track of them after integration. DisGeNET GDA content is represented according to various structured data model conventions: as a relational database, as an RDF linked dataset, and now as a nanopublication linked dataset. In the following sections, we first describe the RDF linked dataset version of DisGeNET data, followed by a description of the DisGeNET nanopublication set and methods to querying them, concluding with possible applications and related work.

Table 1
The DisGeNET evidence classification and its coverage in the nanopublication dataset.

RDF Predicate	ECO Class	DisGeNET Evidence (%)
Assertion Type		
<wi:evidence>	ECO_0000205	CURATED (4)
	ECO_0000266	PREDICTED (1)
	ECO_0000212	LITERATURE (95)
Assertion Method		
<prov:wasGeneratedBy>	ECO_0000218	MANUAL
	ECO_0000203	AUTOMATIC

2.1. RDF Dataset Description

There are three main components in the RDF dataset: GDA content, provenance description of the RDF dataset, and linksets to other linked datasets. Each of these components is described separately in the following sections.

2.1.1. GDA Content

The RDF representation identifies genes by their NCBI Gene ID and diseases by their UMLS Concept Unique Identifier (CUI), and captures the biological type of the association. The gene and the disease additionally have different annotated attributes (see the schema at http://rdf.disgenet.org/). Entities and properties are semantically annotated using standard ontologies such as the NCI thesaurus (NCIt)[5], and resources are identified by using dereferenceable URIs. GDAs are integrated using the DisGeNET association type ontology, which is an ontology developed in the IBI group to fill the gap in formal semantics for the definition of types of associations between a gene and a disease in biological databases. This ontology is based on the description of the GDA association type in the original databases. The DisGeNET ontology is integrated into the Sematicscience Integrated Ontology (SIO) [15], which is an OWL ontology that provides basic types and relations for the description of objects, processes and their attributes, so that GDAs in RDF are semantically harmonized using SIO classes. A normalized dereferenceable URI scheme for the identification of GDA entities is implemented using “http://rdf.disgenet.org/” as namespace.

2.1.2. Provenance

A full provenance description of the RDF linked dataset is provided using the Vocabulary of Interlinked Datasets (VoID) [6], a recommended standard for expressing metadata about RDF datasets. This description includes the provenance of DisGeNET's relational database, its primary databases, and the BeFree text mining approach. The type of curation and level of evidence of each original database is also tracked and annotated. Each data instance in the DisGeNET RDF dataset is explicitly linked to this provenance description in order to granulate and trace back the provenance to the instance level (access to the VoID.ttl file description at http://rdf.disgenet.org/ Web site).

2.1.3. Interlinking

DisGeNET data is linked to LOD in order to both enrich GDAs with annotations from external Semantic Web resources and expand the GDA content in the Semantic Web in a metadata-aware manner. 4,962,315 links to LOD through projects such as Bio2RDF and Linked Life Data, exist in the current version. All entities are linked using the same SKOS[7] predicate *skos:exactMatch*. The interlinking is derived from the cross-references provided by the source databases and ontologies. These cross-references were extracted with in-house scripts and were used for manual interlinking to different data sources. Other linkset statistics between entities can be found on DataHub[8].

[5] http://ncit.nci.nih.gov/

[6] http://www.w3.org/TR/void/

[7] http://www.w3.org/2004/02/skos/

[8] http://datahub.io/es/dataset/disgenet/

3. DisGeNET Nanopublication Dataset

As conforming to the nanopublication standard, our DisGeNET nanopublications consist of four named graphs: head, assertion, provenance and publication information (Figure 1). The head graph defines the structure of the nanopublication by linking to the other nanopublication graph URIs. The assertion graph contains the description for a specific single GDA assertion. The provenance graph includes provenance, evidence and attribution statements that were directly mapped from the VoID description of the RDF dataset. Finally, the publication information graph includes all the metadata information regarding the nanopublication itself. We also include in this graph a description of the general topic of the nanopublication to enhance discoverability. The general topic of our nanopublications is ‘Gene-Disease Association’, and each nanopublication is annotated using the Dublin Core vocabulary (DC)[9] object property ‘subject’ to the SIO concept ‘SIO_000983’, which represents the ‘gene-disease association’ concept. The nanopublication dataset is a linked dataset as all resources are identified and linked to other linked datasets using dereferenceable URIs, and gene and disease entities are identified by URIs constructed on widely adopted namespaces, such as *identifiers.org* [16] to enable interoperability and query federation. Importantly, the nanopublication URIs are *trusty*, i.e. they are generated by applying the Trusty URI approach to guarantee the integrity of the nanopublication content and ensure its immutability, reliability, and verifiability for scientific citation.

3.1. Ontologies

For modeling our nanopublications, we needed to determine what information to include, how to formally represent it, and which ontologies to use to best represent the semantics. To represent the GDAs in the assertion part of the nanopublications, we used the same triples and ontologies already present in the RDF version of our data. That is, we use SIO to encode both the type of association and to relate the disease and the gene associated, and NCIt to encode the gene and disease biomedical entity types.

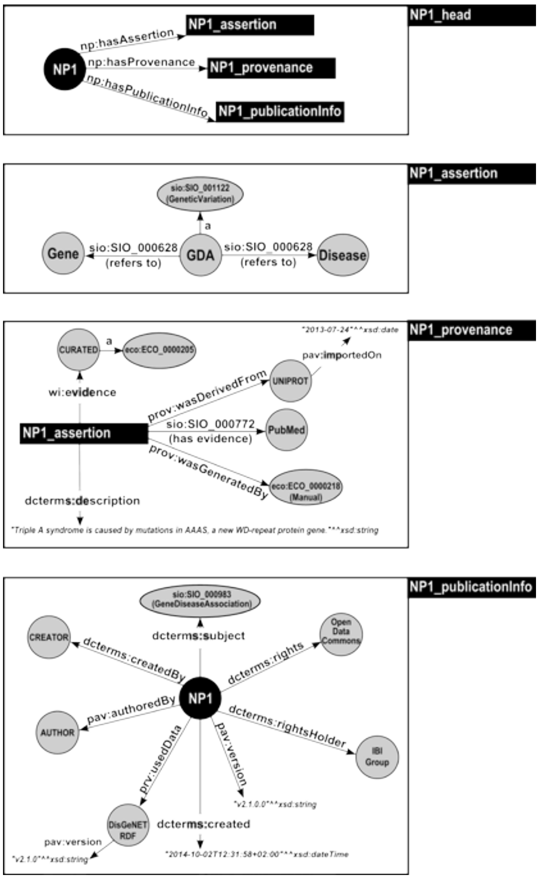


Fig. 1: DisGeNET nanopublication schema. See an example in RDF/TriG notation at our Web site[10].

One important step in the nanopublication modeling was to find appropriate vocabularies for the description of the provenance and metadata in the nanopublication graphs. To represent provenance information we mainly used the PROV Ontology (PROV-O) [11] , which is a W3C recommended standard as a general high-level model for provenance. It provides well-defined upper-level classes, relationships and restrictions to frame any kind of provenance. For authorship, versioning, and content creation we used the Provenance, Authoring and Versioning (PAV) vocabulary [18], and for Web data we used the Provenance Vocabulary Core Ontology Specification (PRV) [12] , which are both extensions of the PROV ontology that allow for more explicit representations of the data. Finally, DC terms are used for general metadata. The evidence

[9] http://dublincore.org/

[10] http://www.disgenet.org/web/DisGeNET/v2.1/rdf#nanoSchema

[11] http://www.w3.org/TR/prov-o/

[12] http://trdf.sourceforge.net/provenance/ns.html

annotation is formally described using the Weighted Interests vocabulary (WI) [13], which provides a suitable object property *wi:evidence* that links the assertion with its evidence, and the Evidence Codes Ontology (ECO)[14], which provides suitable classes to represent the meaning of the DisGeNET evidence classes (Table 1). It is not always straightforward to find ontologies that are semantically compatible with our requirements and existing data, e.g. it was not possible to establish perfect equivalence between our evidence classes and ECO classes. Platforms such as the NCBO Bioportal[15], Ontology Lookup Service[16], or the NCBO Annotator[15] ease this task. Even though the modeling in the Semantic Web offers the freedom to make any decision necessary to best describe the data, it is a good practice to use commonly used vocabularies. With this in mind, we chose the ECO and WI ontologies since they are already in use to model neXtProt evidence [8-9].

A summary of the Linked Data vocabularies used in DisGeNET nanopublications is shown in Table 2. We published our dereferenceable nanopublications on the Web with a human-readable list of the vocabularies. The SIO and ECO ontology concepts are deployed in our triple store to be available both as machine-readable explicitly at axiom level to optimize the GDAs searches in our SPARQL endpoint, and to be human-readable in our Linked Data Faceted Browser.

3.2. Schema

Here we present the first version of the nanopublication model for DisGeNET data, which is based on the official nanopublications guidelines[4] (accessed during summer of 2014). It is worth mentioning that other nanopublications were used as templates (examples accessed at *nanopub.org*, neXtProt nanopublications [8-9], BEL2nanopub approach[17]) to facilitate the process of modeling. Moreover, the reuse of templates fosters a natural creation of best practices both for RDF triple description of information and for vocabularies adoption and, consequently, enhances interoperability of scientific results [11].

The *assertion graph* states the gene and the disease involved in the association, each identified by existing well-defined URIs. It also states the type of GDA as asserted by the authors of the statement. This is annotated using SIO, which it provides formal semantics for representing GDA entities. For example, it is possible to explicitly model the relationships between the gene and the disease in the association as ‘gene-disease association linked with altered gene expression’. The *provenance graph* includes provenance and attribution information directly linked to the assertion such as the scientific article as the primary evidence, the source database from which it was derived and the derivation date for the first stage of annotation. The method of extraction of the assertion is also included, i.e. whether the statement was manually or automatically asserted. The level of evidence of each assertion is annotated using the DisGeNET evidence classification, a classification system used in the context of DisGeNET, similar to those used in other databases such as the quality assessment classification system in neXtProt [8-9]. We minted the DisGeNET evidence concepts using the DisGeNET RDF namespace (http://rdf.disgenet.org) as a base. This is a temporary solution as, in order to make our nanopublications more discoverable for aggregation of evidence, we are planning to use the ConceptWiki [18] to add URLs to the DisGeNET evidence concepts. Additionally, a human readable description of the assertion extracted from the attributed sources is added. Finally, the *publication information graph* includes all the metadata information regarding the nanopublication itself such as when it was made by date/time stamp, copyright information to inform how the nanopublication can be reused, the version of the nanopublication, the RDF Linked Data version used to produce it, and its authors and creators. These attribution statements use object properties referring to unique digital identifiers that represent researchers, e.g. ORCID identifiers.

[13] http://smiy.sourceforge.net/wi/spec/weightedinterests.html

[14] http://www.evidenceontology.org/

[15] http://bioportal.bioontology.org/

[16] http://www.ebi.ac.uk/ontology-lookup/

[17] https://github.com/tkuhn/bel2nanopub/

[18] http://www.conceptwiki.org/

Table 2
Vocabularies prefixes, namespaces and topics used in DisGeNET nanopublications.

Prefix	Namespace	Topic
np	http://www.nanopub.org/nschema#	Nanopublication
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF
xsd	http://www.w3.org/2001/XMLSchema#	XML
sio	http://semanticscience.org/resource/	General science
ncit	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#	Biomedical
lld	http://linkedlifedata.com/resource/umls/id/	Linked Life Data
miriam-gene	http://identifiers.org/ncbigene/	Identifiers.org/NCBI Gene
miriam-pubmed	http://identifiers.org/pubmed/	Identifiers.org/PubMed
eco	http://purl.obolibrary.org/obo/eco.owl#	Evidence codes
wi	http://purl.org/ontology/wi/core#	Preferences in context
prov	http://www.w3.org/ns/prov#	General provenance
pav	http://purl.org/pav/2.0/	Provenance, authoring, versioning
prv	http://purl.org/net/provenance/ns#	Provenance of Web data
dcterms	http://purl.org/dc/terms/	General metadata

3.3. Availability, Production and Sustainability

We present here the first release of DisGeNET published as nanopublications, which corresponds to version v2.1.0.0. The dataset consists of 940,034 nanopublications, representing the same number of scientific assertions for 381,056 different GDAs with their detailed provenance, levels of evidence and publication information descriptions. In total this represents 3,760,136 annotated RDF nanopublication graphs. Specifically, the dataset is composed of 31,961,156 quads, i.e. RDF triples with their graph (or "context") added as the fourth member in the tuple (Subject, Predicate, Object, Context), everything being serialized using the TriG syntax[19]. DisGeNET nanopublications can be accessed in three ways: as a file in TriG format that can be downloaded, by navigating using our Faceted Browser, or by querying our SPARQL endpoint (see http://rdf.disgenet.org/). The dataset is made available under the Open Database License terms whose full text can be found at http://opendatacommons.org/licenses/odbl/1.0/.

The generation of our nanopublications started from the relational database whose data are used to produce the RDF linked dataset. This RDF dataset is generated by in-house scripts that prepare data for the D2RQ platform [20], which is the software that serializes relational data into RDF/Turtle. Nanopublications are derived from the RDF graph linked dataset representation. Both DisGeNET linked datasets are stored and available to query in the Virtuoso [21] SPARQL endpoint. We would like to generate the DisGeNET nanopublications via SPARQL INSERT/CONSTRUCT queries over the RDF dataset, but this is currently not possible for technical limitations of the Virtuoso SPARQL server. To bypass this issue, we developed custom scripts to construct and serialize our nanopublications in the recommended TriG syntax directly from our relational data. While, on the one hand, this solution may be a source of production errors, it may foster a future base of more mature scripts forming a sustainable pipeline to produce DisGeNET nanopublications on a regular basis, and for sharing them with the community through systems as GitHub[22].

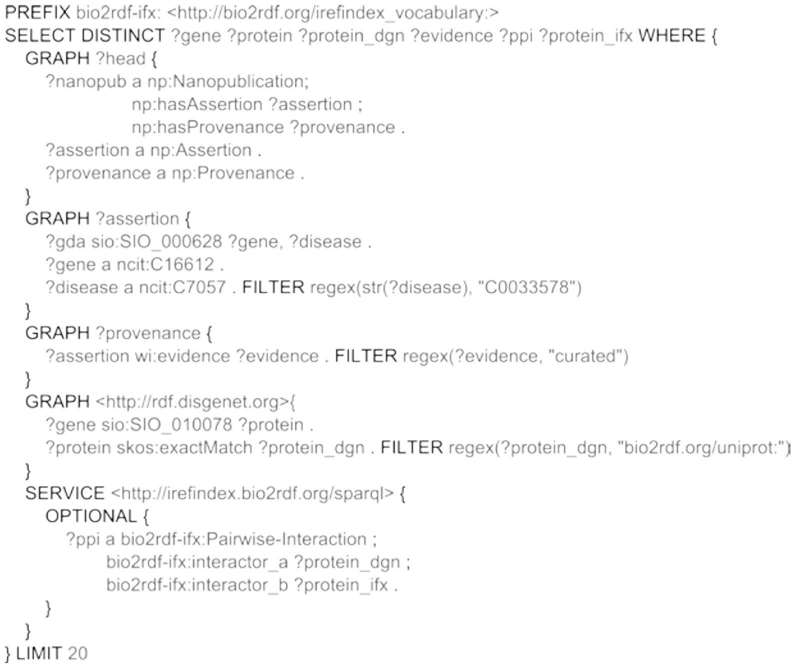
The nanopublication dataset will be updated accordingly in conjunction with its parent relational and RDF versions. Two major updates per year are envisioned for the relational database and consequently the RDF Linked Data distribution, therefore these updates may well also affect the nanopublication content. In addition, the maintenance of the nanopublication dataset may require additional new versions. In each major revision of DisGeNET we include more data sources into the database to increase the coverage on the current knowledge in GDAs, and new annotations, adding more value to the data. For example, in the last update we included new and popular expert-curated datasets: RGD, CTD mouse, CTD rat and our literature-mined BeFree dataset, as well as new annotations describing the level of evidence of each data source. These new evidence codes are highlighted in Table 1. The versioning for nanopublications consists of keeping

[19] http://www.w3.org/TR/trig/
[20] http://d2rq.org/
[21] http://virtuoso.openlinksw.com/
[22] https://github.com/

track of the provenance of both the RDF and the relational version of DisGeNET data from which the RDF is derived. Thus, the nanopublication version information is a composite of: the version of the relational database (v2.1) plus the version of the RDF dataset (v2.1.0) plus the version of the nanopublication (v0). Finally, in recognition of the interest in the nanopublishing of DisGeNET GDAs, we note that from the day we made it available in the download section of our Web site (October 13rd, 2014) until January 26th, 2015, the nanopublication dataset has been downloaded 52 times while the RDF has been downloaded 43 times.

4. Querying DisGeNET Nanopublications

With the aim to show the questions that can be answered by our nanopublication implementation, we use the following question as an example: *What are the proteins (and their protein interactions) associated to Prostatic Neoplasms with curated evidence?* The query in Figure 2 illustrates how to retrieve this information. First, we ask for nanopublications (?nanopub) but only pointing at their assertion and provenance graphs (*Selecting Nanopublication Content by Graph)*. Second, we query DisGeNET for all the genes associated to Prostatic Neoplasms. To retrieve these data we need to query the assertion graph (?assertion), where we ask for gene-disease associations (?gda) where the disease involved in the association is identified by the CUI ‘C0033578’ (*Retrieving Gene-Disease Associations*). Third, we filter the prior results with those assertions annotated as curated in DisGeNET (?evidence). This query part involves the provenance graph (?provenance) (*Filtering By Evidence)*. Finally, we link these results to the Interaction Reference Index database [20], which contains PPI annotations, through the Bio2RDF::irefindex SPARQL endpoint, thereby federating the query (*Linking with Other LOD Resources*). Since our RDF data annotates each gene with the protein(s) that it encodes, we are able to link DisGeNET to Bio2RDF::irefindex by *Protein* resources. To retrieve proteins from DisGeNET we query the RDF graph (http://rdf.disgenet.org). Then, we link these protein resources in DisGeNET to protein resources in Bio2RDF through the corresponding linkset to <http://bio2rdf.org/uniprot:*uniprotID*>. To answer this question, we have retrieved data querying different graphs: head, assertion and provenance graphs in each nanopublication, the DisGeNET-RDF graph, and the iRefIndex graph.



```
PREFIX bio2rdf-ifx: <http://bio2rdf.org/irefindex_vocabulary:>
SELECT DISTINCT ?gene ?protein ?protein_dgn ?evidence ?ppi ?protein_ifx WHERE {
  GRAPH ?head {
    ?nanopub a np:Nanopublication;
            np:hasAssertion ?assertion ;
            np:hasProvenance ?provenance .
    ?assertion a np:Assertion .
    ?provenance a np:Provenance .
  }
  GRAPH ?assertion {
    ?gda sio:SIO_000628 ?gene, ?disease .
    ?gene a ncit:C16612 .
    ?disease a ncit:C7057 . FILTER regex(str(?disease), "C0033578")
  }
  GRAPH ?provenance {
    ?assertion wi:evidence ?evidence . FILTER regex(?evidence, "curated")
  }
  GRAPH <http://rdf.disgenet.org>{
    ?gene sio:SIO_010078 ?protein .
    ?protein skos:exactMatch ?protein_dgn . FILTER regex(?protein_dgn, "bio2rdf.org/uniprot:")
  }
  SERVICE <http://irefindex.bio2rdf.org/sparql> {
    OPTIONAL {
      ?ppi a bio2rdf-ifx:Pairwise-Interaction ;
          bio2rdf-ifx:interactor_a ?protein_dgn ;
          bio2rdf-ifx:interactor_b ?protein_ifx .
    }
  }
} LIMIT 20
```

Fig. 2: An example of SPARQL query.

5. Applications

We aim to incorporate the DisGeNET GDA collection in knowledge discovery projects such as the Open PHACTS Discovery platform [4]. The Open Pharmacological Concepts Triple Store project (Open PHACTS) has developed a powerful cloud-based platform for open access data following a Semantic Web approach that allows scientists to draw on diverse databases to answer many questions relating to drug discovery. The new version of the platform will integrate and provide access to additional datasets such as WikiPathways [22], neXtProt, which was also recently converted into nanopublications, and DisGeNET.

6. Related Work

A variety of datasets represented as nanopublications have recently been published. Beck *et al.* [21] provided a nanopublication dataset on comprehensive genome-wide association studies (GWAS) to organize and annotate the complex spectrum of observed human GWAS phenotypes for reuse and interchange, to assist with cross-species genotype and phenotype comparisons, and to integrate GWAS data into the Linked Data Web. This work underlined the importance of including appropriate provenance and context information to avoid confusion to data consumers since their GWAS nanopublications are simply items of data not yet validated, i.e. not established facts. Chichester *et al.* [9] explored the use of neXtProt nanopublications to obtain new insights based on restricted levels of evidence related to sequence variation, expression, and regulation of human proteins important for precision medicine. Mina *et al.* [11] used the nanopublication model to expose different assertions generated by a Taverna workflow analysis applied to the investigation of the relation between Huntington's Disease genes and epigenetic regulation. They showed the potential of the model to provide metadata from a computational analysis, which will enable reproducibility and increase trust in the assertions. They also showed that the nanopublication model enables the connection of this information to the Research Object model [14]. In Sernadela *et al.* [19], the authors explored the nanopublication integration of large collections of annotated associations between drugs and their adverse events extracted by data mining techniques and applied to pharmacovigilance, and present three interoperable data exchange interfaces. Recently, two novel examples of using nanopublications to track and aggregate evidence for future applications on knowledge discovery and evidence-based decision making processes have appeared. The Repurposing Drugs with Semantics (ReDrugS) framework [12], based on a systems biology approach, represents biological and chemical entity interactions contained in databases as nanopublications including descriptions of the experimental methods used to derive the assertions. By creating consensus assertions, they assign a combined probability of truth inferred from those experimental methods. They showcased their approach by searching and discovering new drug-gene associations for drug repurposing based on statistical aggregation of confidence. Another relevant pioneering work [5] proposes to track the provenance information in diagnostic databases and diagnostic processes by a nanopublication approach, with the goal to enable accurate and evidence-based clinical decision making.

7. Summary

We have created a nanopublication-based linked dataset that provides 940,034 nanopublications on scientific statements of human GDAs. These GDAs identified by Trusty URIs are machine-interpretable, immutable, permanent, and verifiable, which promotes data citations and stable references. Each GDA statement has its provenance description providing attribution, creation time, and further context of its creation to confer trust. Each GDA is classified as "CURATED", "PREDICTED", or "LITERATURE" to categorize the evidence of the statement based on the type of assertion and curation made in the original databases. We have enriched the provenance annotation by stating the type of curation of the assertion, and classified the nanopublications by their level of evidence. DisGeNET nanopublications include metadata annotations about the general topic of the nanopublications, i.e. 'Gene-Disease Association', semantically described by SIO to ease their discoverability in the Semantic Web.

With an illustrative use case we show how our nanopublications can be used to explore GDAs and how they can be integrated with relationships published in other LOD sources, which to permit data integration across domains.

The publication of our DisGeNET nanopublications on the Web of Data will enable a large-scale interconnection of statements about genes and diseases and will allow users to explore them based on evidence. This is essential for knowledge discovery, and our approach can help to get a better picture of the molecular basis of pathological conditions.

Acknowledgments

We thank Dr. Mark Thompson (Leiden University Medical Center) and Dr. Jesse Van Dam (Wageningen UR) for sharing their expertise in the modeling of nanopublications. We also thank the reviewers for their helpful comments. The research leading to these results has received support from the Instituto de Salud Carlos III-Fondo Europeo de Desarollo Regional (PI13/00082 and CP10/00524). Also, from the Innovative Medicines Initiative Joint Undertaking under grants agreements n° 115002 [eTOX] and nº 115191 [Open PHACTS], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. Laura I. Furlong received support from Instituto de Salud Carlos III Fondo Europeo de Desarollo Regional (CP10/00524). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

References

[1] A. Bauer-Mehren, M. Rautschka, F. Sanz, L.I. Furlong, DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks, Bioinformatics 26 (2010), 2924-2926.

[2] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L.I. Furlong, Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research, BMC Bioinformatics 16 (2015), 55.

[3] A. Callahan, J. Cruz-Toledo, P. Ansell, M. Dumontier, Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data, Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, Sebastian Rudolph, The Semantic Web: Semantics and Big Data. 10th International Conference, ESWC 2013, Montpeiller, France, May 26-30, 2013. Proceedings, Lecture Notes in Computer Science. Volume 7882, Springer Berlin Heidelberg, 2013, 200-212.

[4] A. Gray, P. Groth, A. Loizou, S. Askjaer, C. Brenninkmeijer, K. Burger, C. Chichester, C.T. Evelo, C. Goble, L. Harland, S. Pettifer, M. Thompson, A. Waagmeester, A.J. Williams, Applying Linked Data Approaches to Pharmacology: Architectural Decisions and Implementation, Semantic Web Journal 5 (2014), 101-113.

[5] A. Rodríguez-González, M. Martinez-Romero, M. Egaña Aranguren, M.D. Wilkinson, Nanopublishing clinical diagnoses: tracking diagnostic knowledge base content and utilization, 2014 IEEE 27th International Symposium on Computer-Based Medical Systems (CBMS), 2014, 335-340.

[6] B.M. Good, E.L. Clarke, S. Loguercio, A.I. Su, Building a biomedical semantic network in Wikipedia with Semantic Wiki Links, Database (Oxford) (2012), 2012:bar060.

[7] B.M. Good, E.L. Clarke, L. de Alfaro, A.I. Su, The Gene Wiki in 2011: community intelligence applied to human gene annotation, Nucleic Acids Research 40 (2012), D1255-D1261.

[8] C. Chichester, O. Karch, P. Gaudet, L. Lane, B. Mons, A. Bairoch, Converting neXtProt into Linked Data and nanopublications, Semantic Web 6 (2015).

[9] C. Chichester, P. Gaudet, O. Karch, P. Groth, L. Lane, A. Bairoch, B. Mons, A. Loizou, Querying neXtProt nanopublications and their value for insights on sequence variants and tissue expression, Journal of Web Semantics 29 (2014), 15.

[10] D. Rebholz-Schuhmann, A. Oellrich, R. Hoehndorf, Text-mining solutions for biomedical research: enabling integrative biology, Nature Reviews Genetics 13 (2012), 829-839.

[11] E. Mina, M. Thompson, R. Kaliyaperumal, J. Zhao, E. Horst, Z. Tatum, K. Hettne, E.A. Schultes, B. Mons, M. Roos, Nanopublications for exposing experimental data in the life-sciences: a Huntington's Disease case study, Journal of Biomedical Semantics 6 (2015), 5.

[12] J.P. McCusker, R. Yan, K. Solanki, J. Erickson, C. Chang, M. Dumontier, J. Dordick, D. McGuinness, A Nanopublication Framework for Systems Biology and Drug Repurposing, William R. Hogan, Sivaram Arabandi, Mathias Brochhausen, ICBO 2014 International Conference on Biomedical Ontology. Proceedings of the 5th International Conference on Biomedical Ontology. Houston, Texas, USA, October 8-9, 2014, CEUR Workshop Proceedings. Volume 1327, 2014, 90-92.

[13] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, L.I. Furlong, DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes, Database (Oxford) (2015), 2015:bav028.

[14] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garca Cuesta, J.M. Gomez-Perez. G. Klyne. K. Page, M. Roos, J.E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, C. Goble, Workflow-centric research objects: First class citizens in scholarly discourse, Alexander García Castro, Christoph Lange, Frank van Harmelen, Benjamin Good, SePublica 2012 Semantic Publishing. Proceedings of the 2nd Workshop on Semantic Publishing. Hernissos, Crete, Greece, May 28th, 2012, CEUR Workshop Proceedings. Volume 903, 2012, 1-12.

[15] M. Dumontier, C.J.O. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N.R. Del Rio, G. Duck, L.I. Furlong, N. Keath, D. Klassen, J.P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M.D. Wilkinson, R. Hoehndorf, The Semanticscience Integrated

Ontology (SIO) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5 (2014), 14.

[16] N. Juty, N. Le Novère, H. Hermjakob, C. Laibe, Towards the collaborative curation of the registry underlying Identifiers.org, Database (Oxford) (2013), 2013:p.bat017.

[17] N. Queralt-Rosinach and L.I. Furlong, DisGeNET RDF: A Gene-Disease Association Linked Open Data Resource, Adrian Paschke, Albert Burger, Paolo Romano, M. Scott Marshall, Andrea Splendiani, SWAT4LS 2013. Semantic Web Applications and Tools for Life Sciences. Proceedings of the 6th International Workshop on Semantic Web Applications and Tools for Life Sciences. Edinburgh, UK, December 10, 2013, CEUR Workshop Proceedings. Volume 1114, 2013. [DisGeNET RDF Web site at http://rdf.disgenet.org/].

[18] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A.J.G. Gray, C. Goble, T. Clark, PAV ontology: provenance, authoring and versioning, Journal of Biomedical Semantics 4 (2013), 37.

[19] P. Sernadela, P. Lopes, J.L. Oliveira, Exploring nanopublications integration in pharmacovigilance scenarios, 2013 IEEE 15th International Conference on e-Health Networking. Applications and Services (IEEE Healthcom 2013). Lisbon, Portugal, October 9-12th 2013, 2013, 728-730.

[20] S. Razick, G. Magklaras, I.M. Donaldson, iRefIndex: A consolidated protein interaction database with provenance, BMC Bioinformatics 9 (2008), 405.

[21] T. Beck, R.C. Free, G.A. Thorisson, A.J. Brookes, Semantically enabling a genome-wide association study database, Journal of Biomedical Semantics 3 (2012), 9.

[22] T. Kelder, M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C. Evelo, A.R. Pico, WikiPathways: building research communities on biological pathways, Nucleic Acids Research 40 (2012 Database issue), D1301-D1307.

[23] T. Kuhn and M. Dumontier, Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data, Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, Anna Tordai, The Semantic Web: Trends and Challenges. 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings, Lecture Notes in Computer Science. Volume 8465, Springer International Publishing Switzerland, 2014, 395-410.

[24] The UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Research 43 (2015 Database issue), D204-D212.